Mini-Max-Structured Neural Tangent Kernel in Estimating Average Treatment Effect Confounded by Image Co-variate [1][2]

Honor Thesis Defense

Michael Wang

Quantitative Science, Emory University

Table of contents

Introduction

Motivation

- Estimating Average Treatment Effect (ATE) is challenging in observational studies.
- Image co-variates introduce even more complexities if they are confounding variables.
- Some real world examples:
    - medical imaging
    - real estate promotion picture
    - etc.

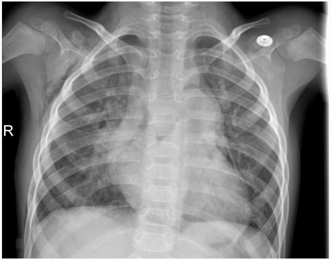
Motivation

We seek co-variate balancing across study groups.

- Inverse Probability Weighting (IPW)?
  - Requires knowledge about confounding co-variates in the estimation process
  - Unstable estimates when extreme propensity score happens
- Introduce a ”better” estimator: Mini-Max-Structured Neural Tangent Kernel:
  - Does not rely on knowledge of co-variates
  - Stability

Semi-Synthetic Data

- In empirical testing stage, we used lung X-ray Pneumonia imaging data [3]
- Treatment and outcome data are unachievable
- Simulation in three frameworks



Semi-Synthetic Data Generation

Frameworks Overview

1. Simple Brightness Framework
2. Label-Based Framework
3. Image Filtering Framework

Framework 1: Brightness Framework

- Average brightness of image:

$$X_i = \frac{1}{224 \times 224} \sum_{i,j} B_{ij}$$

- Propensity score:

$$e(X_i) = \frac{1}{1 + e^{-\alpha(X_i - c)}}$$

- Treatment assignment:

$$W_i \sim \text{Bernoulli}(e(X_i))$$

- Outcome:

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

where $Y_i(1) = 1 + e(X_i)$ and $Y_i(0) = 0 + e(X_i)$

Framework 2: Label-Based Framework

This framework tries to mimic the complex reality, assuming our treatment is antibacterial.

- Co-variate includes both label and brightness:

$$X_i = [L_i, B_i], \quad L_i \in \{\mathrm{NORMAL}, \mathrm{BACTERIA}, \mathrm{VIRUS}\}$$

- Propensity score:

$$\mathrm{logit}(e(X_i)) = \beta_0 + \beta_1 B_i + \beta_2 \mathbb{I}(L_i = \mathrm{BACTERIA}) + \beta_3 \mathbb{I}(L_i = \mathrm{VIRUS})$$

- Treatment assignment:

$$W_i \sim \mathrm{Bernoulli}(e(X_i))$$

- Baseline outcome:

$$\theta(L_i) = \begin{cases} 0 & \text{if } L_i = \mathrm{NORMAL} \\ -1 & \text{if } L_i \in \{\mathrm{BACTERIA}, \mathrm{VIRUS}\} \end{cases}$$

Framework 2: Label-Based Framework

- Treatment effect:

$$\tau(L_i) = \begin{cases} 0 & \text{if } L_i = \text{NORMAL} \\ 1 & \text{if } L_i = \text{BACTERIA} \\ -1 & \text{if } L_i = \text{VIRUS} \end{cases}$$

- Potential outcomes:

$$Y_i(0) = \theta(L_i) + e(X_i), \quad Y_i(1) = Y_i(0) + \tau(L_i)$$

- Observed outcome:

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

Framework 3: Image Filtering Framework

- Image filter matrix:

$$F = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

- Convolution:

$$P'_{ij} = \sum_{u=-1}^{1} \sum_{v=-1}^{1} P_{i+u,j+v} \cdot F_{u+2,v+2}$$

- Filtered brightness (aggregated):

$$X'_i = \frac{1}{H' \cdot W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} (P'_{ij})^2$$

Framework 3: Image Filtering Framework

Apply the *filtered brightness* to Framework 1:

- Propensity score:

$$e(X_i') = \frac{1}{1 + \exp\left(-\alpha(X_i' - c)\right)}$$

- Treatment assignment:

$$W_i \sim \text{Bernoulli}(e(X_i'))$$

- Potential outcomes:

$$Y_i(1) = 1 + e(X_i'), \quad Y_i(0) = 0 + e(X_i')$$

- Observed outcome:

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

Inverse Probability Weighting [4]

IPW 1

Inverse Probability Weighting Estimator 1

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_iY_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)$$

- Goal: Balance co-variates by re-weighting.
- Uses the true propensity score from the data generating process.

IPW 2

Inverse Probability Weighting Estimator 2

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}\right)$$

- Goal: Estimate treatment effect with re-weighted observations.
- $\hat{e}(X_i)$ is estimated using logistic regression on average image brightness X_i.

IPW 3

- Let image i have pixel values $\{P_{i1}, P_{i2}, \ldots, P_{ip}\}$, where $p = 224 \times 224$.
- Fit a Lasso regression:

$$W_i = \alpha + \sum_{j=1}^{p} \beta_j P_{ij} + \epsilon_i, \quad \text{subject to } \sum_j |\beta_j| \leq \lambda$$

- Define pixel-weighted brightness:

$$X_i^* = \sum_{j=1}^{p} \hat{\beta}_j P_{ij}$$

- Estimate $\hat{e}(X_i^*)$ using logistic regression on X_i^*.

Inverse Probability Weighting Estimator 3

$$\hat{\tau}_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i Y_i}{\hat{e}(X_i^*)} - \frac{(1-W_i)Y_i}{1-\hat{e}(X_i^*)}\right)$$

- Goal: Improve propensity score estimation by capturing important pixel-level features.
- $\hat{e}(X_i^*)$ is the estimated propensity score from logistic regression on pixel-weighted brightness.

Mini-max Approach

Mini-max Approach

Bias Formulation

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}\left[\gamma_i f(W_i, X_i) - (f(1, X_i) - f(0, X_i))\right]$$

Where

$$f(W, X) = \beta_0^\top \psi(X)(1 - W) + \beta_1^\top \psi(X) W$$

- We consider a class of functions $f(W, X)$ that describe how outcomes may depend on both treatment and co-variates.
- Here, $\psi(X)$ is a basis function expansion of co-variates (e.g., $[1, X, X^2]$ for quadratic functions)
- Our goal: choose weights γ to minimize the maximum of bias (worst case) over all such functions f.

Mini-max Approach

Apply Cauchy–Schwarz Upper Bound

$$\text{Bias} \leq \|A\gamma - b\|_2 \cdot \left\| \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \right\|_2 \Rightarrow \min_{\gamma} \|A\gamma - b\|_2^2$$

We can then minimize the squared bias upper bound:

$$\hat{\gamma} = \arg\min_{\gamma} \|A\gamma - b\|_2^2 = (A^\top A)^{-1}(A^\top b)$$

Where:

$$A = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \psi(X_i)(1 - W_i) \\ \psi(X_i)W_i \end{bmatrix}, \quad b = \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} -\psi(X_i) \\ \psi(X_i) \end{bmatrix}$$

Regularization term λI added to ensure invertibility:

$$\hat{\gamma} = (A^\top A + \lambda I)^{-1} A^\top b$$

RBF Kernel

What if function f is nonlinear?
Using the Kernel Trick:

$$K(x, x') = \langle \psi(x), \psi(x') \rangle$$

- Replace basis $\psi(X_i)$ with nonlinear kernel similarities by computing inner products.
- Define:

$$K(Z_i, Z_j) = \exp\left(-\frac{\|Z_i - Z_j\|^2}{2\sigma^2}\right)$$

$$Z_i = \begin{bmatrix} \alpha W_i \\ \beta X_i \end{bmatrix} \quad \text{(concatenating treatment + co-variate)}$$

RBF Kernel

Objective Function

$$(K + \lambda I)\gamma = K_{\text{diff}}$$

- Define group-specific similarities:

$$K_1(i) = \sum_j K(Z_i, Z_{1j}), \quad K_0(i) = \sum_j K(Z_i, Z_{0j})$$

$$K_{\text{diff}} = K_1 - K_0$$

- Solve the regularized linear system for $\hat{\gamma}$.

Neural Tangent Kernel Definition

$$K\left(\begin{bmatrix} W \\ X \end{bmatrix}, \begin{bmatrix} W' \\ X' \end{bmatrix}\right) = \langle \nabla_\theta \hat{f}(X), \nabla_\theta \hat{f}(X') \rangle$$

$f(X)$: output of a neural network with input image X and parameters θ.

Neural Network Modeling:

- Train CNN separately on treated and control groups.
- Compute group-specific gradients:

$$f_1(X) : \text{trained on treated}, \quad f_0(X) : \text{trained on controlled}$$

Treatment-Specific Kernel Construction

$$K\left(\begin{bmatrix} W \\ X \end{bmatrix}, \begin{bmatrix} W' \\ X' \end{bmatrix}\right) = \begin{cases} \langle \nabla_\theta \hat{f}_1(X), \nabla_\theta \hat{f}_1(X') \rangle, & \text{if } W = W' = 1 \\ \langle \nabla_\theta \hat{f}_0(X), \nabla_\theta \hat{f}_0(X') \rangle, & \text{if } W = W' = 0 \\ 0, & \text{if } W \neq W' \end{cases}$$

Define counterfactual similarity vectors:

$$K_{\text{diff},0}(i) = \sum_j \nabla_\theta \hat{f}_0(X_i)^\top \nabla_\theta \hat{f}_0(X_j), \quad \text{if } W_i = 0$$

$$K_{\text{diff},1}(i) = \sum_j \nabla_\theta \hat{f}_1(X_i)^\top \nabla_\theta \hat{f}_1(X_j), \quad \text{if } W_i = 1$$

Solve the system for weights:

$$(K + \lambda I)\gamma = K_{\text{diff}}$$

Augmented Inverse Probability Weighting

Augmented Inverse Probability Weighting (AIPW)

Utilizing the group-specific models trained, we can further calculate Augmented Inverse Probability Weighting (AIPW) estimator:

AIPW Estimator

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_1(X_i) - \hat{f}_0(X_i)\right\} + \frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_i \cdot \left\{Y_i - \hat{f}_{W_i}(X_i)\right\}$$

with

$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\hat{\gamma}_i^2\left(Y_i - \hat{f}(X_i)\right)^2$$

- $\hat{f}_1(X_i), \hat{f}_0(X_i)$: predicted outcomes from treatment-specific models.
- $\hat{\gamma}_i$: balancing weights from IPW, Minimax, or kernel estimator.
- The first term is a regression estimator (model-based), and the second term is a bias correction via weighting.

Oracle and Pixel-Based Estimators

Oracle Estimators

An idealized estimator that assumes knowledge of the underlying data-generating process.

- IPW 1 using the true propensity score, $e(X_i)$
- IPW 2 using estimated propensity score, $\hat{e}(X_i)$, by logistic regression
- IPW with Linear Mini-Max Approach
- IPW with RBF Kernel Mini-Max Approach

Pixel-Based Estimators

Pixel-Based Estimators, without assumes the knowledge of parameters and co-variate structures, relies purely on pixel values from the image.

- IPW 3 using Lasso-regressed weighted brightness to estimate propensity score by logistic regression $\hat{e}(X_i^*)$
- IPW with NTK Mini-Max Approach
- AIPW with NTK Mini-Max Approach

Empirical Results

Results

Method	$\hat{\tau}_0$	Sample σ	Coverage	$E_n[\hat{\tau}_i]$
Truth	1	NA	NA	NA
IPW I	0.863	0.197	0.93	1.043
IPW II	0.898	0.283	0.99	1.065
IPW III	0.765	0.072	0.71	0.893
IPW w/ Linear Mini-Max	1.019	0.011	0.94	0.993
IPW w/ RBF Mini-Max	0.994	0.011	0.9	0.992
IPW w/ NTK Mini-Max	1.103	0.032	0.29	1.091
AIPW w/ NTK Mini-Max	1.011	0.015	0.66	1.022

Table 1: Semi-Synthetic Framework 1

Results

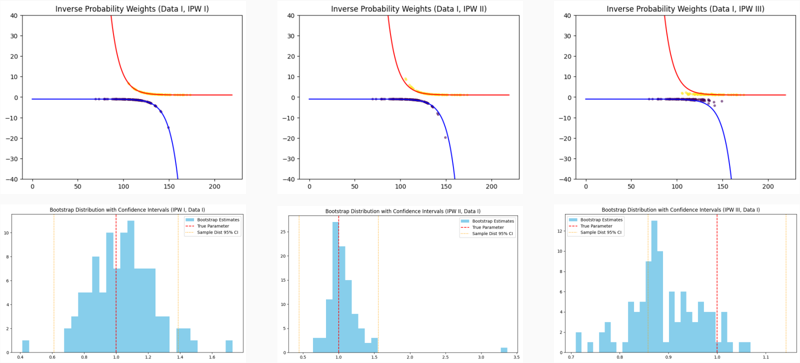


Figure 1: Semi-Synthetic Framework 1

Results

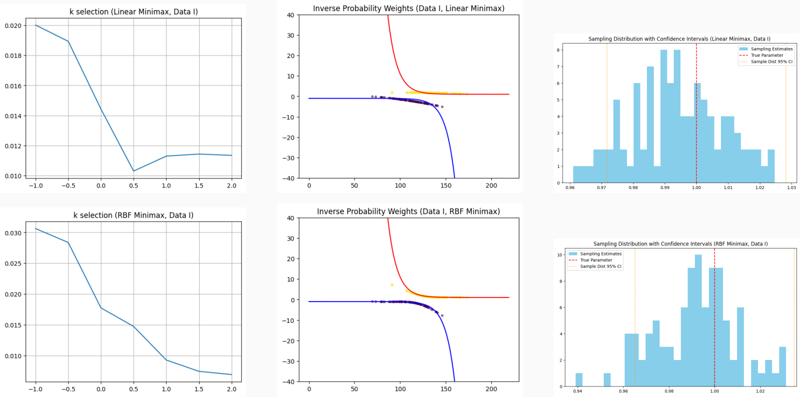


Figure 2: Semi-Synthetic Framework 1

Note: selecting $\lambda = \frac{1}{n^k}$, $k = 0$, $n = 200$, $\lambda = 1$

Results

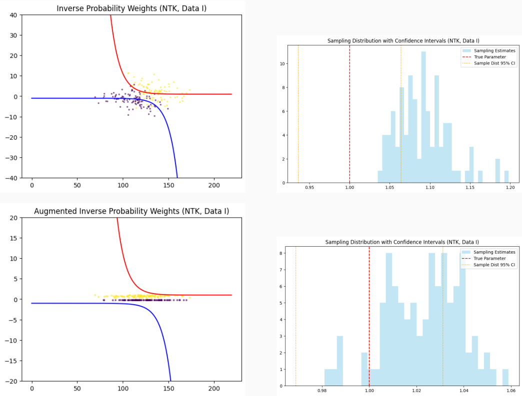


Figure 3: Semi-Synthetic Framework 1: NTK IPW (top 2) & AIPW (bottom 2)

Note: λ *selection:* $\lambda_{NTK_{IPW}} = 2.5e^{-5}, \lambda_{NTK_{AIPW}} = 9000$

Results

Method	$\hat{\tau}_0$	Sample σ	Coverage	$E_n[\hat{\tau}_i]$
Truth	0.227	NA	NA	NA
IPW I	0.362	0.137	0.95	0.254
IPW II	0.337	0.116	0.96	0.242
IPW III	0.312	0.068	0.76	0.306
IPW w/ Linear Mini-Max	0.437	0.063	0.95	0.245
IPW w/ RBF Mini-Max	0.339	0.088	0.97	0.249
IPW w/ NTK Mini-Max	0.303	0.136	0.86	0.331
AIPW w/ NTK Mini-Max	0.761	0.035	0	0.747

Table 2: Semi-Synthetic Framework 2

Results

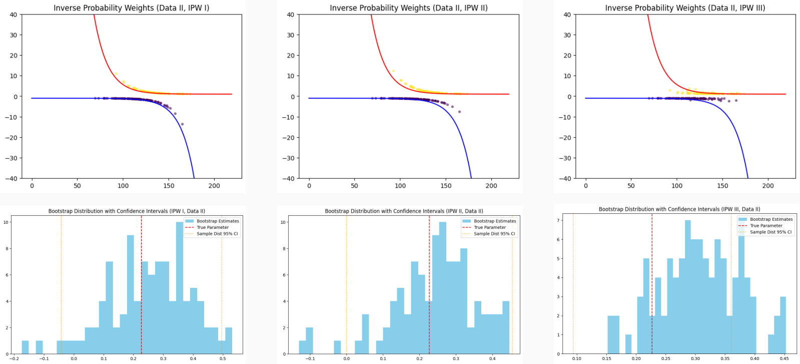


Figure 4: Semi-Synthetic Framework 2

Results

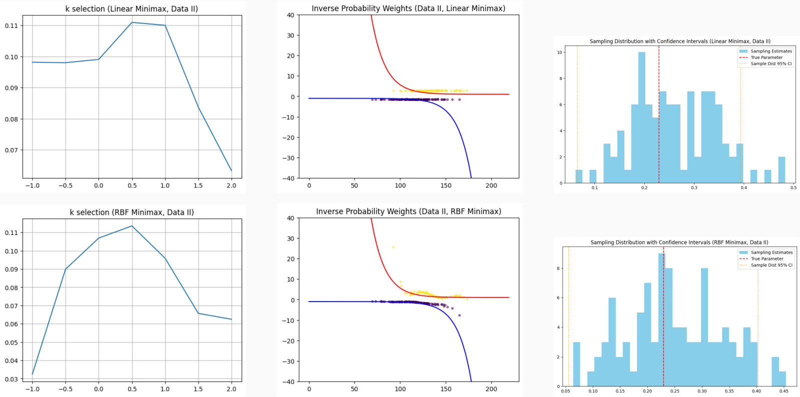


Figure 5: Semi-Synthetic Framework 2

Note: λ *selection:* $k = 2$, $n = 200$, $\lambda_{linear} = \frac{1}{n^k} = 2.5e^{-5}$,
$k = 1$, $n = 200$, $\lambda_{rbf} = \frac{1}{n^k} = 0.005$

Results

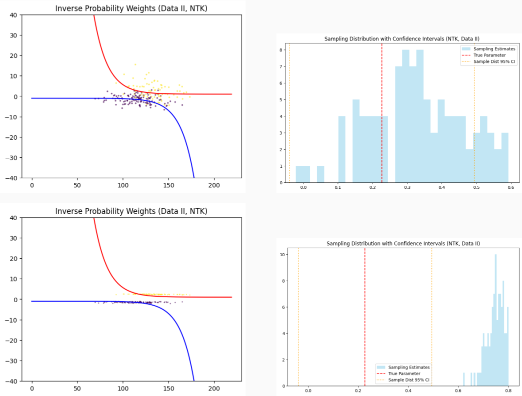


Figure 6: Semi-Synthetic Framework 2: NTK IPW (top 2) & AIPW (bottom 2)

Note: λ *selection:* $\lambda_{NTK_{IPW}} = 0.001$, $\lambda_{NTK_{AIPW}} = 90$

Results

Method	$\hat{\tau}_0$	**Sample** σ	**Coverage**	$E_n[\hat{\tau}_i]$
Truth	1	NA	NA	NA
IPW I	1.006	0.195	0.94	1.008
IPW II	0.997	0.091	0.95	1.023
IPW III	0.812	0.073	0.44	0.846
IPW w/ Linear Mini-Max	0.971	0.015	0.52	0.968
IPW w/ RBF Mini-Max	1.004	0.014	0.82	1.016
IPW w/ NTK Mini-Max	1.138	0.044	0.05	1.162
AIPW w/ NTK Mini-Max	1.013	0.004	0.52	1.007

Table 3: Semi-Synthetic Framework 3

Results

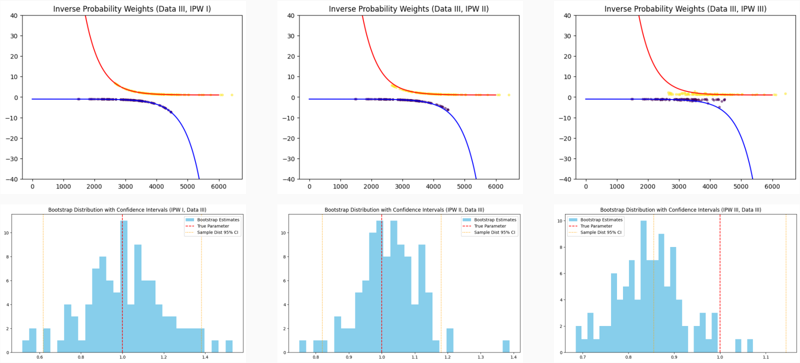


Figure 7: Semi-Synthetic Framework 3

Results

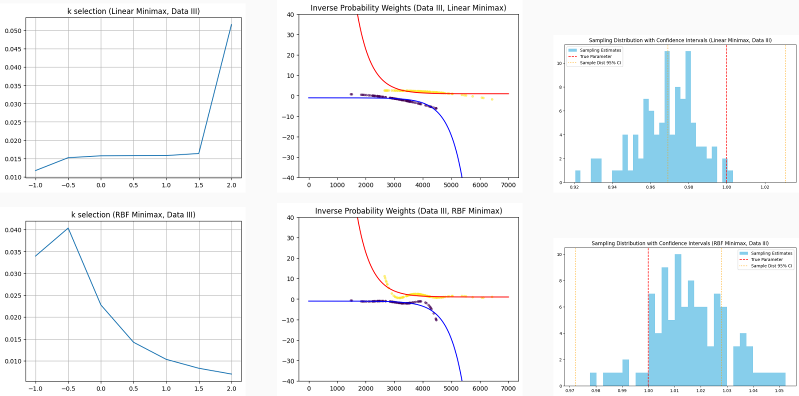


Figure 8: Semi-Synthetic Framework 3

Note: selecting $\lambda = \frac{1}{n^k}$, $k = 0.5$, $n = 200$, $\lambda = 0.0707$

Results

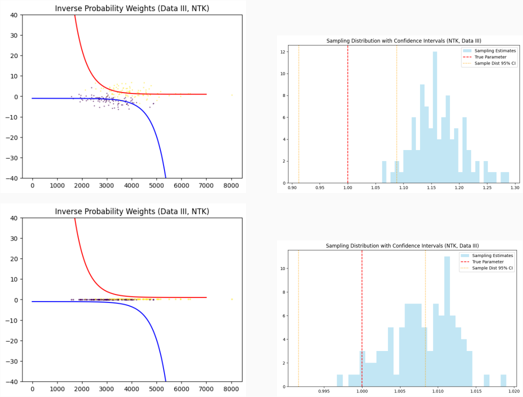


Figure 9: Semi-Synthetic Framework 3: NTK IPW (top 2) & AIPW (bottom 2)

Note: λ *selection:* $\lambda_{NTK_{IPW}} = 0.005$, $\lambda_{NTK_{AIPW}} = 24000$

Discussion and Conclusion

Discussion

- **Oracle and Purely Pixel-Based:** The Oracle Estimators generally performs well in the measuring results provided. Our proposed Mini-Max Structured Neural Tangent Kernel outperforms IPW III in most cases in terms of bias and standard error, although the purely pixel-based estimators do not perform as well as the oracle estimators.
- **Strength:** The NTK-based estimator is able capture nonlinear, high-dimensional image structure without prior knowledge about feature information.
- **Limitations:** Despite strong point estimates, NTK IPW performs badly in coverage estimated by the sampling distribution.

Discussion

- **AIPW:** AIPW with NTK improved some estimates but suffered in coverage, especially in Framework 2. The augmentation was sensitive to model training quality and the possibility of overfitting.
- **Variance and Large Weights:** NTK might predict large (positive or negative) weights, rendering the estimation highly unstable.
- **Regularization:** The regularization term is huge for AIPW, while the estimated balancing weights are still bad.

Conclusion

This work proposed a **Mini-Max-Structured Estimation Framework** that uses Neural Tangent Kernels to estimate ATE from image-confounded data. It **does not require prior knowledge of confounding co-variates** and uses only pixel-level information for balancing. Among all estimators tested:

- **Oracle Estimators** served as a useful benchmark and reinforced the benefit of structural information of features.
- **Pixel-Based estimators** showed strong potential but require more work on fine tuning and cross validation.

Future Work

- **Improved Regularization Tuning:** Explore better cross-validation mechanism for automatic selection of λ to stabilize NTK-based estimates.
- **Neural Network Model:** Better Neural Network Model would produce a more robust gradient estimation of parameters in NTK, which helps obtain a better result.
- **Real-World Applications:** Apply to observational medical image datasets where expert annotations or external instruments are available for validation.

Q & A

Eli Ben-Michael, Avi Feller, David A. Hirshberg, and José R. Zubizarreta.
The balancing act in causal inference, 2021.

David A. Hirshberg, Arian Maleki, and Jose R. Zubizarreta.
Minimax linear estimation of the retargeted mean, 2021.

Daniel Kermany, Kang Zhang, and Michael Goldbaum.
Large dataset of labeled optical coherence tomography (oct) and chest x-ray images, 2018.

PAUL R. ROSENBAUM and DONALD B. RUBIN.
The central role of the propensity score in observational studies for causal effects.
Biometrika, 70(1):41–55, 04 1983.